PCT/EP2004/051784

BUNDESREPUBLIK DEUTSCHLAND

EP04/51784

REC'D 04 NOV 2004

WIPO PCT

Prioritätsbescheinigung über die Einreichung einer Patentanmeldung

Aktenzeichen:	103 37 823.5
Anmeldetag:	18. August 2003
Anmelder/Inhaber:	Siemens Aktiengesellschaft, 80333 München/DE
Bezeichnung:	Sprachsteuerung von Audio- und Videogeräten
IPC:	G 10 L, G 06 F

Die angehefteten Stücke sind eine richtige und genaue Wiedergabe der ursprünglichen Unterlagen dieser Patentanmeldung.

München, den 23. September 2004
Deutsches Patent- und Markenamt
Der Präsident
Im Auftrag

Wallner

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

A 9161
06/00
EDV-L

Beschreibung

Sprachsteuerung von Audio- und Videogeräten

Durch die Gesetzgebung und zur Erhöhung der Sicherheit wird die Benutzung von Spracherkennung bei Applikationen im Automotive-Bereich in Zukunft verstärkt Anwendung finden. Neben Telephonieanwendungen werden Sprachsteuerungen mittlerweile auch für Telematiksysteme, Infotainment-Systeme und In-Car-Systeme wie Klimaanlagen eingesetzt. Das verwendete Vokabular ist durch die aktuellen Erkenner bedingt einfach strukturiert und in der Regel kommandobasiert.

Die Sprachsteuerung von CD-Geräten erfolgt dabei in derzeitigen Produkten mittels Kommandos für die Grundbefehle wie etwa "Stopp", "Play", "Pause" etc. Die Auswahl der zu spielenden Titel wird mittels der Nummer des Titels eingegeben, also beispielsweise durch "Play 5". Der Erkenner kann sich dabei auf das Erkennen des Kommandowortes in Verbindung mit einer Zahl beschränken. Da der Benutzer jedoch die Zuordnung der Titel zur Nummer auf der CD oftmals nicht kennt, ist dies Vorgehensweise unkomfortabel.

Davon ausgehend liegt der Erfindung die Aufgabe zugrunde, die Bedienung von Audio- und Videogeräten einfacher, komfortabler und sicherer zu machen.

Diese Aufgabe wird durch die in den unabhängigen Patentansprüchen angegebenen Erfindungen gelöst. Vorteilhafte Ausgestaltungen ergeben sich aus den abhängigen Ansprüchen.

Dementsprechend sind in einem Verfahren zur Spracherkennung Audiodaten und/oder Videodaten jeweils Textdaten zugeordnet. In einer Graphem-zu-Phonem-Konvertierung werden die Textdaten als Grapheme Phonemen zugeordnet. Sodann können die Textdaten mit ihren zugehörigen Phonemen als Vokabular eines Spracherkenners verwendet werden.

Dadurch ergibt sich ein sehr reduziertes und auf die jeweilige Audio- und/oder Videoanwendung spezifiziertes Erkennervokabular, das auch von einem Spracherkenner mit sehr geringen Ressourcen verarbeitet werden kann, wie er üblicherweise bei eingebetteten Spracherkennungslösungen im Auto oder in anderen Video- und/oder Audiogeräten vorliegt.

Durch diese Vorgehensweise wird es ermöglicht, einen Titel beispielsweise durch "Play Waterloo" oder nur "Waterloo" direkt einzugeben, ohne dass der Benutzer sich während der Autofahrt nebenbei noch die richtige Titelnummer überlegen muss. Speziell bei Audiosystemen mit CD-Wechslern ist ein direkter Zugriff wünschenswert.

Vorzugsweise liegen die Audiodaten auf einer CD vor. Soweit die CD CD-Text aufweist, sind die den Audiodaten zugeordneten Textdaten auf der CD als CD-Text gespeichert. Diese können dann direkt für die Graphem-zu-Phonem-Konvertierung herangezogen werden.

Die den Audio- und/oder Videodaten zugeordneten Textdaten können auch allgemein in einem Inhaltsverzeichnis des Speichermediums gespeichert sein, das die Audio- und/oder Videodaten enthält.

Die Audiodaten können beispielsweise MP3-Daten sein. Dann sind die Textdaten vorzugsweise in einer Playlist gespeichert.

Alternativ oder ergänzend können die den Video- und/oder Audiodaten zugeordneten Textdaten von einer zentralen Datenbank abgerufen werden, insbesondere über das Internet aus einer Internet-Datenbank.

Die Textdaten enthalten vorzugsweise den Namen des oder der Interpreten und/oder den Titel der Audio- und/oder Videodaten, denen sie zugeordnet sind.

In einem weiteren Schritt können die Textdaten über eine Text-zu-Sprache-Konvertierung akustisch ausgegeben werden, so dass dem Benutzer seine Wahlmöglichkeiten, insbesondere hinsichtlich Titel und Interpreten, vorgelesen werden.

Insbesondere wird über das Verfahren ein Audio- und/oder Videogerät mit Hilfe des Spracherkenners gesteuert.

Eine Anordnung, die eingerichtet ist, eines der geschilderten Verfahren auszuführen, lässt sich zum Beispiel durch Programmieren und Einrichtung einer Datenverarbeitungsanlage mit zu den genannten Verfahrensschritten gehörigen Mitteln realisieren.

Die Anordnung kann beispielsweise ein Autoradio, insbesondere integriert mit Navigationssystem, ein CD-Spieler und/oder ein DVD-Spieler sein.

Weitere Merkmale und Vorteile der Erfindung ergeben sich aus der Beschreibung von Ausführungsbeispielen.

In einem Verfahren zur Spracherkennung wird eine Graphem-zu-Phonem-Technologie bei einem eingebetteten Spracherkenner dazu verwendet, dass die Titelnamen von Songs in Phonem-Folgen konvertiert werden und als Erkennervokabular zur sprachlichen Ansteuerung von CD-, DVD- und/oder MP3-Playern eingesetzt werden. Dies erlaubt dem Benutzer eine direkte Anwahl der Songs über Titel, Interpreten oder alternativ konventionell über die gewohnte Nummern-Nomenklatur.

Werden zu den als Vokabular aufbereiteten Titeln verschiedener CDs die zugeordneten Positionen im CD-Wechsler vermerkt, kann bei der sprachlichen Eingabe der Titel erkannt

und einer bestimmten CD zugeordnet werden. Der Wechsler kann die gewünschte CD einlegen und den gewählten Song abspielen. Die Vokabulargröße in einem 5-fach-Wechsler mit jeweils 20 Liedern pro CD beträgt demnach ca. 100 Einträge. Dies stellt eine Vokabulargröße dar, die mit gängiger Technologie von eingebetteten Spracherkennern abgedeckt werden kann.

Da Song-Titel in unterschiedlichen Sprachen vorliegen können, ist vor der Konvertierung der Titel in Phonem-Folgen eine Sprachidentifikation durchzuführen, die den geeigneten Phonem-Set und die korrekten sprachspezifischen Konvertierungsregeln festlegt.

Bei Audio-CDs liegen die Song-Titel in Textform auf CD-Text-kompatiblen CDs vor. Als alternative Lösung in vernetzten Fahrzeugen kann die Titelliste per Download zur Verfügung gestellt werden.

Es werden also Textdaten von Audio- und/oder Videomedien als Vokabularbasis für den Spracherkenner verwendet. Die direkte Sprachanwahl von Songtiteln erlaubt eine komfortable und für den Fahrer wenig ablenkende Methode zur Bedienung des CD- und MP3-Equipments in Fahrzeugen. Durch die Nutzung der Graphem-zu-Phonem-Technologie kann diese direkte Sprachanwahl realisiert werden und dem Benutzer im Rahmen seiner Sprach-Bedienoberfläche zur Verfügung gestellt werden.

Das vorgestellte Verfahren ist aufgrund seiner Sichtbarkeit an der Benutzeroberfläche leicht nachweisbar. Durch die deutliche Komforterhöhung ist der Mehrwert durch den Benutzer groß und erkennbar. Da sich sprecherunabhängige Systeme auf längere Frist auch im Automotive-Bereich durchsetzen werden, bietet sich eine sprachliche CD- und/oder DVD-Ansteuerung als ideale Ergänzung an.

Das Verfahren kann beispielsweise direkt für CDs im CD-Text-Format verwendet werden. Auf einer Audio-CD sind neben den

eigentlichen Musikdaten noch Zusatzdaten gespeichert, so genannte "Sub-Channels". Es gibt dabei acht Sub-Channels (p, q, r, s, t, u, v und w). Der q-Sub-Channel enthält beispielsweise Informationen über die momentane Position. Eine besondere Stellung nimmt der Leadin-Bereich ein. Der Leadin-Bereich ist ein Bereich vor den normalen Musikdaten und enthält in den q-Sub-Channels die "Table of Contents" (TOC) der CD, also das Inhaltsverzeichnis der CD. In der TOC sind die Anfangspositionen der einzelnen Tracks gespeichert. In den Sub-Channels r-w des Leadins werden nun die CD-Text-Informationen gespeichert, beispielsweise der Name der CD, die Namen der Tracks sowie die Interpreten.

Mit dieser Information kann für den Spracherkenner ein Vokabular dynamisch erzeugt werden. Dank Graphem-zu-Phonem-Konvertierung können dabei die Textdaten in Erkenner-verständliche Phonemketten umgesetzt werden. Zur Bedienung kann dann das Vokabular oder Teile davon zur Steuerung des Audio- und/oder Videogeräts verwendet werden.

Patentansprüche

1. Verfahren zur Spracherkennung,
- bei dem Audiodaten jeweils Textdaten zugeordnet sind,
- bei dem Grapheme der Textdaten Phonemen zugeordnet werden,
- bei dem die Textdaten mit ihren zugehörigen Phonemen als Vokabular eines Spracherkenners verwendet werden.

2. Verfahren nach Anspruch 1,
bei dem die Audiodaten auf einer CD vorliegen.

3. Verfahren nach Anspruch 2,
bei dem die den Audiodaten zugeordneten Textdaten auf der CD als CD-Text gespeichert sind.

4. Verfahren nach einem der vorhergehenden Ansprüche,
bei dem die Audiodaten MP3-Daten sind.

5. Verfahren nach Anspruch 4,
bei dem die Textdaten in einer Playlist gespeichert sind.

6. Verfahren zur Spracherkennung,
- bei dem Videodaten jeweils Textdaten zugeordnet sind,
- bei dem die Textdaten als Grapheme Phonemen zugeordnet werden,
- bei dem die Textdaten mit ihren zugehörigen Phonemen als Vokabular eines Spracherkenners verwendet werden.

7. Verfahren nach einem der vorhergehenden Ansprüche,
bei dem die Textdaten in einem Inhaltsverzeichnis auf einem Speichermedium gespeichert sind, auf dem die Audio- und/oder Videodaten gespeichert sind.

8. Verfahren nach einem der vorhergehenden Ansprüche,
bei dem die Textdaten von einer zentralen Datenbank abgerufen werden, insbesondere über das Internet.

9. Verfahren nach einem der vorhergehenden Ansprüche, bei dem die Textdaten den Namen des Interpreten und/oder den Titel der Audio- und/oder Videodaten enthalten, denen sie zugeordnet sind.

10. Verfahren nach einem der vorhergehenden Ansprüche, bei dem ein Audio- und/oder Videogerät über den Spracherkenner gesteuert wird.

11. Verfahren nach einem der vorhergehenden Ansprüche, bei dem die Textdaten zumindest teilweise in einer Text-zu-Sprache-Konvertierung konvertiert und akustisch ausgegeben werden.

12. Anordnung, die eingerichtet ist, ein Verfahren nach zumindest einem der vorstehenden Ansprüche auszuführen.

13. Anordnung nach Anspruch 12, dadurch gekennzeichnet, dass die Anordnung ein Auto, ein Autoradio, ein CD-Spieler und/oder ein DVD-Spieler ist.

Zusammenfassung

Sprachsteuerung von Audio- und Videogeräten

Zu Audio- und/oder Videodaten vorliegende Textinformationen werden in einer Graphem-zu-Phonem-Konvertierung Phonemen zugeordnet und als Vokabular für einen Spracherkenner verwendet.